DARPA's Big Mechanism Program

Paul R. Cohen

November 20, 2014

When it comes to complicated systems such as the climate, and economic, ecological and biological systems, our influence runs ahead of our understanding. This can be perilous. The Defense Advanced Research Projects Agency (DARPA) created the Big Mechanism program (BMP) to develop technology for constructing, understanding and reasoning about big, complicated systems. The focus of the program is cancer signaling pathways, but the technology is intended to apply more generally. This paper describes the scientific and technical challenges of the BMP.

Imagine a future in which computers read the literature of cancer biology, extract fragments of causal mechanisms from publications, assemble these mechanisms into executable models of unprecedented scale and fidelity, use these models to explain and predict aspects of cancer biology, and even test these predictions *in vitro*. The future is not yet here (the BMP started in July, 2014) but we can see its outlines and our first steps are clear. Roughly half of the program's effort goes into machine reading, the other half into assembling large, causal models and reasoning with them. The BMP makes good use of existing online sources of biological knowledge, as well as existing information extraction methods; while it develops representations and inference methods for causal biological models, methods for machines to understand the claims and evidence in papers, algorithms for inferring causal relationships given data, and robotic techniques for testing computer-generated hypotheses. The BMP focuses on Ras-driven cancers [28, 33, 59], as these are among the least well-understood and most lethal cancers.

1 What is a Big Mechanism?

Science progresses from descriptions of specific phenomena to general, explanatory theories. When explanatory theories are elusive, predictive models may still be useful. Big mechanisms are predictive models or explanatory theories that involve very large numbers of causal relations. This is an idealization, of course, because not all the relations in big mechanisms will be causal; and some relations that are thought to be causal are later shown not to be so; and as a practical matter, scientists will often base predictions on non-causal correlates of causal factors. Still, big mechanisms differ from the associative models of data mining to the extent that they capture causal relationships.

Signaling pathways are a good domain for the BMP because signaling reduces to biochemistry and physics, and what could be more causal than that? But not all physicochemical models are the same: Systems biologists make tradeoffs between the completeness of models, the uncertainty associated with the parameters of these models, the amount of data available for estimating these

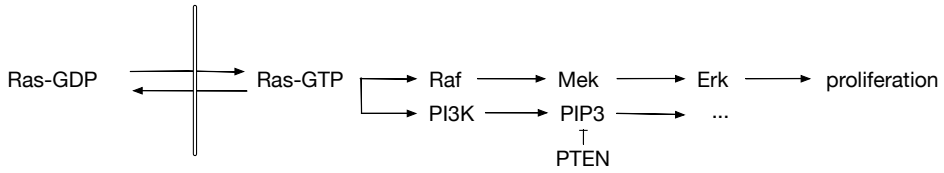

Figure 1: An incomplete signaling pathway map of Ras (excerpted from [59])

parameters, the "readability" of a model by other biologists, and the tractability of computation with the models (see [1, 27, 58, 4] for good introductions to these issues). The biologists in the BMP generally favor rule-based and agent-based simulation models or logical rewriting models that capture some but not all of the details of the underlying biochemistry [4, 14, 21, 43, 61, 62]. These approximations are necessary because of the combinatorial complexity of molecular interactions and the lack of empirical data to calibrate model parameters for these interactions. Also, there is some concern that model parameters do not constrain model behavior [23]. So for the time being, big mechanism models of signaling will be "coarse" in various ways. The challenge will be to manage multiple drafts of models in a way that makes them amenable to revision by both humans and machines [41, 45, 44, 16].

The BMP intentionally does not commit to a single standard for formal biological models. At present, BMP researchers are working with BioPAX [16], BEL [8], Pathway Logic (PL) [61, 62], BioNetGen [20] and Kappa [14] representations, and some researchers are developing high-level programming languages that compile down to ODE models and other formalisms [4, 21, 44]. Ideally, the BMP will produce causal, executable models [26, 27, 45] that make strong predictions through a combination of logical or probabilistic reasoning and powerful simulations.

Pathway "maps" – like the one shown in Figure 1, only much bigger – fall short of this ideal, but can still be useful. Pathway maps are graphs in which the vertices are named entities (typically proteins) and the edges usually represent protein-protein interactions. They are the weakest of models because they make few if any testable predictions. They are generally incomplete and they tell us nothing about the dynamic behavior of the cell over time.[1] That said, pathway maps summarize some of what we know about signaling; they are the starting point for more expressive graphical representations and interactomes and scaffolds [16, 32, 37, 42, 57, 51]; and they are an appropriate target representation for machine reading, which at present extracts little more than interactions between proteins from text.

In addition to models of biological processes, the BMP is developing novel representations for the contents of tables, for the structure of experimental protocols [52], for typical statistical analyses, for rhetorical devices and epistemic or qualified assertions [17, 35, 56], and for other kinds of contents of research papers. These kinds of content are often "meta" to the biological models, but they are important elements of publications because they establish the reasons to believe or doubt models. If a paper says, "mutant and wildtype responded differently to x, suggesting a therapeutic effect of x on y," then a Big Mechanism system should extract both a model fragment that relates x and y and also the uncertainty conveyed by the "suggesting" construction.

[1] Frank McCormick, personal communication

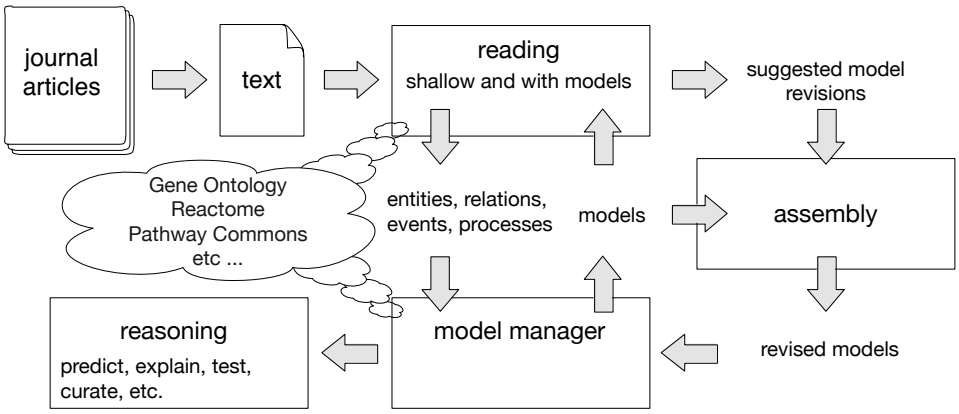

Figure 2: A rough architecture for Big Mechanism systems

2 Big Mechanism Systems

Several consortia of academic and industry researchers are building Big Mechanism systems[2]. The technologies and use cases for these systems vary, but all of the systems are being developed along the lines shown in Figure 2. Big Mechanism systems are expected to read as researchers do: with models in mind. A system will first do shallow reading to discover the entities, relations, events and processes that a text describes. Shallow reading will help a model manager retrieve one or more extant models that the text might be discussing. Deep reading, also called *reading with a model,* will discover what the text says about these models.

One might think that machines would have little difficulty reading papers in cancer biology because the language they contain seems very "close" to biological phenomena. To a human, a sentence like, "Ras acts as a molecular switch that is activated upon GTP loading and deactivated upon hydrolysis of GTP to GDP" seems clear enough, but consider what a machine must do to understand it (see also [30]):

- Figure out the entities – Ras, GTP and GDP – and associate them with standard ontology entries when possible;
- Understand coreferences, or multiple references to the same entity (e.g., the second mention of GTP refers to the same entity as the first one) and anaphoric references (e.g., "that" refers to "molecular switch," which itself refers to Ras);
- Recognize activation, loading, and hydrolysis as events;
- Correctly place the entities in their appropriate roles in these events; for example, the subject of activation is the molecular switch, and the subject of loading is not GTP but the switch;
- Weave these events into a causal story that unfolds over time. In this case, time is ordinal and events are treated as instantaneous (indicated by "upon"), but in general, sentences may contain kinematic information and finer temporal distinctions.

[2]The participants are: University of Arizona (Surdeanu, PI); Carnegie Mellon University (Hovy, PI and Miskov-Zivanov, PI); University of Chicago (Rzhetsky, PI); Cycorp, (Lenat and Witbrock, PIs); Harvard University (Fontana, PI and Sorger, PI); IHMC (Allen, PI); USC/ISI (Marcu, PI); Leidos (Morgenstern and Demir, PIs); SIFT (Burstein, PI); SRI International (Freitag and Rohwer, PIs). MITRE (Hirschman, PI) will conduct evaluations.

"Nicotine induces the binding of Raf-1 to Rb in HAEC and A549 cells as seen by IP/Western blotting."	statement1 = bind(entity1=Raf1,entity2=Rb) statement2 = induce(entity1=Nicotine, process1 = statement1)
"Rb binds to E2F" "transactivation capacity of E2F" "Rb ... inhibits the transactivation capacity of E2F."	ComplexAssembly(In(Rb, E2F), Out(Complex(Rb, E2F))) TemplateRegulation(In(E2F)) DownRegulation(Controller(Rb), Controlled(TemplateRegulation(In(E2F))))
"MEK2/ERK2 catalyzes its own phosphorylation."	(c / catalyze-01 :ARG0 (m / macro-molecular-complex :part (e / enzyme :name (n / name :op1 "MEK2")) :part (e2 / enzyme :name (n2 / name :op1 "ERK2"))) :ARG1 (p / phosphorylate-01 :ARG1 m))

Table 1: Some intermediate representations of sentences.

The Big Mechanism consortia are adapting prior information extraction and machine reading methods to deal with cancer biology texts. In most cases, the approach is to create an intermediate formal representation that "connects" a text to biological models in the sense that it can be derived from text and its elements can be unified with elements of biological models (e.g., [2, 5, 13, 31, 43, 49, 54]). As can be seen in Table 1, these representations are syntactically diverse, but all are hierarchical and composed of predicate-argument assertions, where the predicate is an event such as binding or a modifier such as concentration and the arguments are generally biological entities[3]. Extracting these structures from texts generally requires rules that look for patterns in text (e.g., equations 2-9 of [31]), although sometimes the rules work on the results of parsing the text. These rules are built by humans, or machines can learn them from annotated corpora (e.g., [11, 34, 65]).

Current approaches can extract entities from text and find mentions of events and processes with middling accuracy (e.g., [34, 39, 40, 65, 66, 50]). The consortia will try to improve on these technologies by reading with models, that is, exploiting the context provided by models. For example, if one has a prior model of molecular switches that includes "slots" for activating and de-activating events, and the model specifies that the same entity can participate in both activating and de-activating the switch, then a likely interpretation of the two mentions of GTP is that they are coreferring: GTP is one entity in both the activating and de-activating events.

Barely six months into the BMP, we see only limited attempts at reading with models, and much of that is concerned with coreference resolution and its cousin, event coreference resolution [29]. But models do far more than help to resolve ambiguities in text. For example, Emek Demir and the Leidos Big Mechanism team realized tables encode relationships among the biological entities represented by columns, inducing a kind of model. Moreover, each row represents an instance of each of these relationships. For example, some tables encode triples of the form `columnName(rowName,cellEntry)`, where the column name is a relationship such as `phosporylates` and the row name and cell entry are the respective subject and object (or vice versa) of the relationship. The precise nature of the relationships encoded in a table may not be clear, especially because column and row headings may be arbitrary symbols. One way to determine

[3]The three rows of the table are based on personal communications from Andrey Rzhetsky, Mihai Surdeanu, and Daniel Marcu, respectively.

their meaning is to ask whether whether any row in the table corresponds to a known, standardized biological model. This is accomplished by searching BioPAX for fragments that contain the terms in the row labels and cell entries. If some models are found, they are processed to construct a template model that covers most or all of them. Assuming that all the rows in the table have identical semantics, each row is an instance of the template model, and so can used to enrich BioPAX. Thus, a few rows serve to retrieve model fragments and construct a model, and the remaining rows – which may number in the hundreds – constitute new model instances.

The focus on the relationships between models and text (including tables) is a distinctive aspect of the program, one that is expected to extract much more content from the literature than current methods do. The relationships between text and models that matter most, of course, suggest *revisions* to models. A text might suggest *specializing* part of a model by, for example, introducing a novel protein species. Or a text might *narrow the scope* of a model to a particular tissue type. A text might *generalize* a mechanism, or suggest that several mechanisms are actually just one, parameterized differently. Texts can *suggest alternative mechanisms*, such as alternative signaling cascades that target the same cell function. Some texts address models indirectly, by speaking to the evidence that supports the models; for example, texts can contradict previous findings, thereby *eroding support* for extant models.

The results of reading, then, are suggested revisions to prior models and the evidence that supports them. The task of evaluating these suggestions and implementing some of them is called assembly. At present, no Big Mechanism system does assembly, although most participants in the BMP have developed technology to organize large numbers of logical or probabilistic assertions into theories, or consistent sets, or most-probable groupings.

The greatest challenge posed by assembly will be managing the proliferation of models and metadata that stand in various relations to each other (the model manager box in Fig. 2). It is unrealistic to think that there is "one true model" of anything and that the job of Big Mechanism systems is to discover it. Even if there were one true model of, say, one particular Ras-driven cancer, it would not spring into being fully formed like Athena from the brow of Zeus, but would be uncovered in stages, by consulting multiple publications, and it would be revised in light of new evidence. How are multiple drafts of models maintained? How shall we organize the evidence for these models? How shall we recognize the overlapping or contradicting parts of models [18, 38]? How shall we organize models of "one thing" – say, Ras signaling – that are actually families of models specific to organisms, tissue types, and mutations? The complexity of model management is compounded by the need for qualitatively different kinds of models for different kinds of reasoning. As noted, pathway maps might serve to summarize molecular interactions, while physicochemical models – more or less coarse-grained – are required to simulate the kinetics of these interactions.

Abstraction, modularity and re-use of models will probably facilitate model management [44], although to the best of our knowledge, model management by machines is a new area of science.

Any solution to the model-management problem should depend on the anticipated uses of models (see the box labeled "reasoning" in Fig. 2). The Big Mechanism consortia are focusing on several use cases, of which the most promising are curation and predicting and testing the effects of interventions. Curation refers to a growing movement in biology to have humans translate scientific articles into formal representations that can be processed by machines [7, 9, 24]. Even if Big Mechanism technologies do not automate curation, they could reduce the human effort expended on mundane aspects of the job [25, 49, 55]. And humans could help solve problems that the machines

may not be able to solve alone, such as figuring out which assertions in papers are relevant to a user's curation task [50], or detecting subtle elisions or obfuscations that are intended to make results seem more general or convincing. Machines might have trouble understanding things that are not said or are said artfully, and might do well to rely on humans to uncover rhetorical subtleties. One particularly attractive aspect of human-computer curation is that each chunk of work provides training data for machine learning, hastening the day when machines can curate all by themselves.

The second major use case for the BMP is testing interventions. Several BMP teams are focusing on predicting effects of interventions and testing these predictions in simulations that run on conventional computers as well as on Field Programmable Gate Array platforms [47]. The BMP is also able to test predictions *in vitro*, including a fully-automated robotic system that can set up experiments to test hypotheses generated by Big Mechanism systems [36, 46]. Note that making predictions about interventions does not require complete physicochemical models with correct kinetics and all their parameters correctly specified. This is a good thing, not only because these models are hard to build and test [27] but also because of a concern that model parameters do not constrain model behavior [23].

This use case stops short of drug discovery. The BMP focuses on computer-based technologies for understanding complicated systems, and to demonstrate understanding it suffices to make correct predictions about interventions. It is not necessary to go further and actually develop the interventions into drugs, nor does the BMP have the expertise to do so.

3 Big Mechanism and the Practice of Science

Every data point is a complicated function of causal influences. The methods of experimental science give us knowledge of two kinds: They establish causality among tiny numbers of factors, and they quantify the influence of other factors as variance. But of course, this is an idealization. p values generally are pointless, few people report effect sizes, and experiments are done in dramatically different contexts. All of which means that the literature comprises thousands of *narrow* causal assertions of varying quality. We try to figure out complicated systems assertion by assertion. No wonder replicability can be a problem.

However, the purpose of the BMP is not to overturn the experimental methods that have given us great scientific insights. In fact, the BMP assumes that, for now, humans are the best practitioners of experimental methods, and takes researchers at their word until there's a reason not to. This is one reason why the BMP begins with research papers and tries to assemble models that integrate many causal assertions, *collectively*. This approach addresses implicitly the vulnerability of taking authors at their word. Often, the assertions of authors matter less than how these assertions stand in relation to others. Often, these assertions are inconsistent or mutually contradictory [18, 38]. The better we can assemble causal assertions from many sources into large models, the better we will be able to evaluate new assertions. Of course, "we" includes computers, as only computers can evaluate how a causal assertion stands in relation to thousands of others derived from an enormous literature.

With the exception of some work on algorithms for inferring causal structure from data, and the possibility that data from BMP-proposed experiments will feed back into modeling, the BMP does nothing with primary data. Its "data" are the assertions made by authors. There is plenty of

work on inducing biological models directly from data (e.g., [6, 48, 51, 60, 64]), and clearly the bewildering diversity of oncogenic mutations recommends big data approaches to therapy [63]. But the BMP is not a big data program, a cancer therapy program, or even a cancer biology program. It is a program designed to develop technologies to help scientists understand very complicated systems, generally.

The biggest contribution of Big Mechanism might be to change how knowledge is organized and communicated. We still work in a medieval tradition in which each of us searches the literature for very large numbers of narrow results and pulls them into our heads for synthesis into big mechanisms or causal understandings. This "pull" scholarship is enormously wasteful. It works less and less well because there are simply too many results to synthesize in our heads. We can't read any faster than we already do, but the literature grows at an increasing rate, so we read more narrowly. Just when we need to understand highly connected systems *as systems* our research methods force us to focus on little parts. If the BMP works, it will be the first demonstration of a new kind of "push" scholarship where, instead of pulling results into our heads, we push results into machine-maintained big mechanisms, where they can be examined by anyone. This could change science profoundly.

References

[1] Bree B. Aldridge, John M. Burke, Douglas A. Lauffenburger and Peter K. Sorger. Physicochemical modelling of cell signalling pathways. Nature Cell Biology 8, 1195 - 1203 (2006) doi:10.1038/ncb1497 http://www.nature.com/ncb/journal/v8/n11/full/ncb1497.html

[2] Allen, J., M. Swift, et al. (2008). Deep Semantic Analysis of Text. Symposium on Semantics in Systems for Text Processing (STEP). Venice, Italy. http://www.cs.rochester.edu/research/cisd/pubs/2008/step08.pdf

[3] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. Nature genetics, 25(1), 25-29.

[4] John Bachman and Peter Sorger. 2013. New approaches to modeling complex biochemistry. Mol Syst Biol. 2013; 9: 646. Published online Feb 19, 2013. doi: 10.1038/msb.2013.1 http://www.nature.com/nmeth/journal/v8/n2/full/nmeth0211-130.html

[5] Laura Banarescu , Claire Bonial , Shu Cai , Madalina Georgescu , Kira Griffitt , Ulf Hermjakob , Kevin Knight , Philipp Koehn , Martha Palmer , Nathan Schneider (2013). Abstract Meaning Representation for Sembanking. Proc. ACL Linguistic Annotation Workshop http://amr.isi.edu/

[6] Gurkan Bebek and Jiong Yang. 2007. PathFinder: mining signal transduction pathway segments from protein-protein interaction networks. BMC Bioinformatics 2007, 8:335 doi:10.1186/1471-2105-8-335

[7] Bourne PE, McEntyre J (2006) Biocurators: Contributors to the World of Science. PLoS Comput Biol 2(10): e142. doi:10.1371/journal.pcbi.0020142

[8] http://www.openbel.org

[9] Sarah Burge, Teresa K. Attwood, Alex Bateman, Tanya Z. Berardini, Michael Cherry, Claire O'Donovan, loannis Xenarios, and Pascale Gaudet Biocurators and Biocuration: surveying the 21st century challenges Database 2012: bar059 doi:10.1093/database/bar059 published online March 20, 2012

[10] Cerami, E.G., Gross, B.E., Demir, E.,; Rodchenkov, I., Babur, O., & Anwar, N. (2011): Pathway Commons, a web resource for biological pathway data. In Nucleic acids research 39 (Database issue), pp. D685-90.

[11] K. B. Cohen, K. Verspoor, M. Bada, M. Palmer, and L. Hunter, The Colorado Richly Annotated Full Text (CRAFT) corpus: Multi-model annotation in the biomedical domain, In press.

[12] D. Croft, G. O'Kelly, G. Wu, R. Haw, and M. Gillespie, et al., Reactome: A Database of Reactions, Pathways and Biological Processes, Nucleic Acids Res 39: D691?7, doi:10.1093/nar/gkq1018, (2011).

[13] http://www.cyc.com/cyc/cycl/syntax

[14] V. Danos, J. Feret, W. Fontana, R. Harmer and J. Krivine. "Rule-Based Modeling of Cellular Signaling" Concurrency Theory 2007, Lecture Notes in Computer Science, 4703:17-41

[15] Danos V, Feret J, Fontana W, Harmer R, Krivine J (2009) Rule-based modelling and model perturbation. Lect Notes Comput Sci 5750: 116?137

[16] Emek Demir, Michael P Cary, Suzanne Paley, Ken Fukuda, Christian Lemer, Imre Vastrik, Guanming Wu, Peter D'Eustachio, Carl Schaefer, Joanne Luciano, Frank Schacherer, Irma Martinez-Flores, Zhenjun Hu, Veronica Jimenez-Jacinto, Geeta Joshi-Tope, Kumaran Kandasamy, Alejandra C Lopez-Fuentes, Huaiyu Mi, Elgar Pichler, Igor Rodchenkov, Andrea Splendiani, Sasha Tkachev, Jeremy Zucker, Gopal Gopinath, Harsha Rajasimha, Ranjani Ramakrishnan, Imran Shah, Mustafa Syed, Nadia Anwar, zgn Babur, Michael Blinov, Erik Brauner, Dan Corwin, Sylva Donaldson, Frank Gibbons, Robert Goldberg, Peter Hornbeck, Augustin Luna, Peter Murray-Rust, Eric Neumann, Oliver Ruebenacker, Matthias Samwald, Martijn van Iersel, Sarala Wimalaratne, Keith Allen, Burk Braun, Michelle Whirl-Carrillo, Kei-Hoi Cheung, Kam Dahlquist, Andrew Finney, Marc Gillespie, Elizabeth Glass, Li Gong, Robin Haw, Michael Honig, Olivier Hubaut, David Kane, Shiva Krupa, Martina Kutmon, Julie Leonard, Debbie Marks, David Merberg, Victoria Petri, Alex Pico, Dean Ravenscroft, Liya Ren, Nigam Shah, Margot Sunshine, Rebecca Tang, Ryan Whaley, Stan Letovksy, Kenneth H Buetow, Andrey Rzhetsky, Vincent Schachter, Bruno S Sobral, Ugur Dogrusoz, Shannon McWeeney, Mirit Aladjem, Ewan Birney, Julio Collado-Vides, Susumu Goto, Michael Hucka, Nicolas Le Novre, Natalia Maltsev, Akhilesh Pandey, Paul Thomas, Edgar Wingender, Peter D Karp, Chris Sander and Gary D Bader. 2010. Nat Biotechnol 28 (9), pp. 935–942.

[17] A de Waard, HP Maat. 2012. Epistemic modality and knowledge attribution in scientific discourse: A taxonomy of types and overview of features Proceedings of the Workshop on Detecting Structure in Scholarly Discourse, 47-55 `http://www.aclweb.org/anthology/W/W12/W12-4306.pdf`

[18] Divoli, A., E.A. Mendonca, J.A. Evans, and A. Rzhetsky. 2011. Conflicting biomedical assumptions for mathematical modeling: the case of cancer metastasis. PLoS Computational Biolology 7(10). doi: 10.1371/journal.pcbi.1002132.

[19] Dogrusoz, U., Cetintas, A., Demir, E., & Babur, O. (2009). Algorithms for effective querying of compound graph-based pathway databases. BMC bioinformatics, 10(1), 376.

[20] Faeder JR, Blinov ML, Hlavacek WS (2009) Rule-based modeling of biochemical systems with BioNetGen. Methods Mol Biol 500: 113?167

[21] Feret J, Danos V, Krivine J, Harmer R, Fontana W. 2009. Internal coarse-graining of molecular systems. Proc Natl Acad Sci USA 106: 6453?6458 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2672529/?report=reader#!po=90.0000

[22] Friedman N. 2004. Inferring cellular networks using probabilistic graphical models. Science. 2004 Feb 6;303(5659):799-805

[23] Gutenkunst R.N., Waterfall J.J., Casey F.P., Brown K.S., Myers C.R., Sethna J.P. 2007. Universally sloppy parameter sensitivities in systems biology models. PLoS Comput. Biol. 2007;3:1871?1878. doi:10.1371/journal.pcbi.0030189 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2000971/?report=reader

[24] Jodi Hirschman,Tanya Z. Berardini, Harold J. Drabkin, and Doug Howe. 2010. A MOD(ern) perspective on literature curation Mol Genet Genomics. May 2010; 283(5): 415?425. Published online Mar 11, 2010. doi: 10.1007/s00438-010-0525-8

[25] Lynette Hirschman, Gully A. P. C Burns, Martin Krallinger, Cecilia Arighi, K. Bretonnel Cohen, Alfonso Valencia, Cathy H. Wu, Andrew Chatr-Aryamontri, Karen G. Dowell, Eva Huala, Anlia Loureno, Robert Nash, Anne-Lise Veuthey, Thomas Wiegers, and Andrew G. Winter. 2012. Text mining for the biocuration workflow. Database Vol. 2012, bas020; doi:10.1093/database/bas020

[26] W. S. Hlavacek, J. R. Faeder, M. L. Blinov, R. G. Posner, M. Hucka and W. Fontana. 2006. "Rules for Modeling Signal-Transduction Systems" Science STKE, Vol. 2006, Issue 344, pp. re6, 18 July 2006

[27] W. S. Hlavacek. How to deal with large models. Mol. Syst. Biol. 2009; 5: 240. Published online Jan 20, 2009. doi: 10.1038/msb.2008.80 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2644176/?report=reader#!po=91.6667

[28] Matthew Holderfield, Marian M. Deuker, Frank McCormick and Martin McMahon. 2014. Targeting RAF kinases for cancer therapy: BRAF-mutated melanoma and beyond. Nature Reviews Cancer 14, 455?467 (2014) doi:10.1038/nrc3760 Published online 24 June 2014

[29] Hovy, E.H., T. Mitamura, M.F. Verdejo, J. Araki, A. Philpot. 2013. Events are Not Simple: Identity, Non-Identity, and Quasi-Identity. Proceedings of the 1st Events Workshop at the conference of the HLT-NAACL. Atlanta, GA

[30] Lawrence Hunter and K. Bretonnel Cohen. Biomedical Language Processing: Perspective What's Beyond PubMed? Mol Cell. Mar 3, 2006; 21(5): 589-594. doi: 10.1016/j.molcel.2006.02.012 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1702322/

[31] Hunter, L., Lu, Z., Firby, J., Jr., W. B., Johnson, H., Ogren, P., & Cohen, K. (2008). OpenDMAP: An open source, ontology-driven concept analysis engine with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. BMC Bioinformatics, 9:78. http://www.ncbi.nlm.nih.gov/pmc/articles/

PMC2275248/#S2

[32] Ideker T, Lauffenburger D. 2003. Building with a scaffold: emerging strategies for high- to low-level cellular modeling. Trends Biotechnol. 2003 Jun;21(6):255-62.

[33] Karnoub, A. E., & Weinberg, R. A. (2008). Ras oncogenes: split personalities. Nature reviews Molecular cell biology, 9(7), 517-531.

[34] Kim, J.-D., Pyysalo, S., Ohta, T., Bossy, R., Nguyen, N. and Tsujii, J.. (2011). Overview of BioNLP Shared Task 2011. In: Proceedings of the BioNLP Shared Task 2011 Workshop, pp. 1–6, Association for Computational Linguistics

[35] Kim, S.M. and E.H. Hovy. 2006. Identifying and Analyzing Judgment Opinions. Proceedings of the Human Language Technology / North American Association of Computational Linguistics conference (HLT-NAACL 2006).

[36] Ross D. King, Jem Rowland, Stephen G. Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N. Soldatova, Andrew Sparkes, Kenneth E. Whelan, and Amanda Clare. The automation of science. Science 3 April 2009: 324 (5923), 85-89. [DOI:10.1126/science.1165620]

[37] Hiroaki Kitano, Akira Funahashi, Yukiko Matsuoka and Kanae Oda Using process diagrams for the graphical representation of biological networks. Nature Biotechnology 23, 961 - 966 (2005) Published online: 4 August 2005 doi:10.1038/nbt1111 http://www.nature.com/nbt/journal/v23/n8/full/nbt1111.html

[38] Kirouac, D.C., Saez-Rodriguez, J., Swantek, J., Burke, J., Lauffenburger, D.A., & Sorger, P.K. (2012): Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks. In BMC systems biology 6 (1), p. 29.

[39] Martin Krallinger, Alfonso Valencia and Lynette Hirschman. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biology 2008, 9(Suppl 2):S8 doi:10.1186/gb-2008-9-s2-s8 http://genomebiology.com/2008/9/S2/S8

[40] Krallinger, M., Vazquez, M., Leitner, F., Salgado, D., Chatraryamontri, A., Winter, A., & Valencia, A. (2011). The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. BMC bioinformatics, 12(Suppl 8), S3.

[41] Le Novere, Nicolas, Finney, Andrew, Hucka, Michael, Bhalla, Upinder S, Campagne, Fabien, Collado-Vides, Julio, Crampin, Edmund J, Halstead, Matt, Klipp, Edda, Mendes, Pedro, Nielsen, Poul , Sauro, Herbert, Shapiro, Bruce, Snoep, Jacky, Spence, Hugh and Wanner, Barry. 2005. Minimum information requested in the annotation of biochemical models (MIRIAM). Nature Biotech, 23:12, 1509–1515, http://dx.doi.org/10.1038/nbt1156

[42] Le Novere, N., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., & Kitano, H. (2009). The systems biology graphical notation. Nature biotechnology, 27(8), 735-741.

[43] Patrick. D. Lincoln and Carolyn Talcott, Symbolic Systems Biology And Pathway Logic, in Sriram Iyengar, Ed., Symbolic Systems Biology. Jones and Bartlett, 2010.

[44] Lopez CF, Muhlich JL, Bachman JA, Sorger PK. 2013. Programming biological models in Python using PySB. Mol Syst Biol. 2013; 9: 646. Published online Feb 19, 2013. doi:

10.1038/msb.2013.1. http://msb.embopress.org/content/9/1/646.long

[45] Aneil Mallavarapu, Matthew Thomson, Benjamin Ullian, and Jeremy Gunawardena. Programming with models: modularity and abstraction provide powerful capabilities for systems biology. J R Soc Interface. Mar 6, 2009; 6(32): 257?270. Published online Jul 22, 2008. doi: 10.1098/rsif.2008.0205 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2659579/?report=reader#!po=2.27273

[46] Yihui Liu; Martin, K.; Sparkes, A.; King, R.D., The Analysis of Yeast Cell Morphology Using a Robot Scientist, Computational Intelligence and Security (CIS), 2010 International Conference on , vol., no., pp.10,14, 11-14 Dec. 2010 doi: 10.1109/CIS.2010.10

[47] N. Miskov-Zivanov, A. Bresticker, D. Krishnaswamy, S. Venkatakrishnan, D. Marculescu, and J. R. Faeder. 2011. Emulation of biological networks in reconfigurable hardware. In Proc. Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine, 2011, pp. 536-540.

[48] Mitra, Koyel and Carvunis, Anne-Ruxandra and Ramesh, Sanath Kumar and Ideker, Trey. 2013. Integrative approaches for finding modular structure in biological networks. Nature Reviews Genetics 14, 719?732 (2013) doi:10.1038/nrg3552 http://www.nature.com/nrg/journal/v14/n10/full/nrg3552.html#ref2

[49] Rafal Rak, Andrew Rowley, William Black, and Sophia Ananiadou Argo: an integrative, interactive, text mining-based workbench supporting curation. Database 2012: bas010 doi:10.1093/database/bas010 published online March 20, 2012

[50] K. E. Ravikumar, K. B. Wagholikar, and Hongfang Liu. 2014. Challenges in adapting text mining for full text articles to assist pathway curation. In Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '14). ACM, New York, NY, USA, 551-558. DOI=10.1145/2649387.2649444 http://dl.acm.org/citation.cfm?doid=2649387.2649444

[51] Rhodes, Daniel R and Tomlins, Scott A and Varambally, Sooryanarayana and Mahavisno, Vasudeva and Barrette, Terrence and Kalyana-Sundaram, Shanker and Ghosh, Debashis and Pandey, Akhilesh and Chinnaiyan, Arul M. 2005. Probabilistic model of the human protein-protein interaction network. Nature Biotechnology 23, 951 - 959 (2005) doi:10.1038/nbt1103 http://www.nature.com/nbt/journal/v23/n8/pdf/nbt1103.pdf

[52] Thomas A Russ, Cartic Ramakrishnan, Eduard H Hovy, Mihail Bota and Gully APC Burns. 2011. Knowledge engineering tools for reasoning with scientific observations and interpretations: a neural connectivity use case. BMC Bioinformatics 2011, 12:351 doi:10.1186/1471-2105-12-351. http://www.biomedcentral.com/1471-2105/12/351

[53] Rzhetsky A, Zheng T, Weinreb C (2006) Self-Correcting Maps of Molecular Pathways. PLoS ONE 1(1): e61. doi:10.1371/journal.pone.0000061

[54] Rzhetsky A, Koike T, Kalachikov S, Gomez SM, Krauthammer M, Kaplan SH, Kra P, Russo JJ, Friedman C. A knowledge model for analysis and simulation of regulatory networks. Bioinformatics. 2000 Dec;16(12):1120-8. http://bioinformatics.oxfordjournals.org/content/16/12/1120.long

[55] Rodriguez-Esteban R, Iossifov I, Rzhetsky A (2006) Imitating Manual Curation of Text-Mined

Facts in Biomedicine. PLoS Comput Biol 2(9): e118. doi:10.1371/journal.pcbi.0020118

[56] Agnes Sandor and Anita de Waard. 2012. Identifying Claimed Knowledge Updates in Biomedical Research Articles. Proceedings of the Workshop on Detecting Structure in Scholarly Discourse, 47-55 http://www.aclweb.org/anthology/W/W12/W12-4302.pdf

[57] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. Nov 2003; 13(11): 2498?2504.doi: 10.1101/gr.1239303 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC403769/#!po=3.57143

[58] Michael W Sneddon, James R Faeder and Thierry Emonet. 2011. Efficient modeling, simulation and coarse-graining of biological complexity with NFsim. Nature Methods 8, 177?183 (2011) doi:10.1038/nmeth.1546 http://www.nature.com/nmeth/journal/v8/n2/full/nmeth.1546.html

[59] Andrew G. Stephen, Dominic Esposito, Rachel K. Bagni, Frank McCormick. Dragging Ras Back in the Ring. Cancer Cell. Volume 25, Issue 3, p272?281, 17 March 2014. http://dx.doi.org/10.1016/j.ccr.2014.02.017

[60] A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science 10 October 2003: 302 (5643), 249-255.Published online 21 August 2003 [DOI:10.1126/science.1087447]

[61] Carolyn Talcott, Steven Eker, Merrill Knapp, Patrick Lincoln, Keith Laderoute. 2004. Pathway Logic Modeling of Protein Functional Domains in Signal Transduction. From Proceedings of the Pacific Symposium on Biocomputing.

[62] Carolyn Talcott. 2006. Formal Executable Models of Cell Signaling Primitives by in 2nd International Symposium On Leveraging Applications of Formal Methods, Verification and Validation ISOLA06

[63] J. M. Tenenbaum and J. Shrager, Cancer: A computational disease that AI can cure, AI Magazine, vol. 32, no. 2, pp. 14?26, 2011.

[64] Guanming Wu, Xin Feng and Lincoln Stein. 2010. A human functional protein interaction network and its application to cancer data analysis. Genome Biology 2010, 11:R53 http://genomebiology.com/2010/11/5/R53

[65] Cecilia N Arighi, Zhiyong Lu, Martin Krallinger, Kevin B Cohen, W John Wilbur, Alfonso Valencia, Lynette Hirschman and Cathy H Wu. 2011. Overview of the BioCreative III Workshop. BMC Bioinformatics 2011, 12(Suppl 8):S1 doi:10.1186/1471-2105-12-S8-S1

[66] http://2013.bionlp-st.org/tasks/pathway-curation